Nurse *et al. Secur Inform (2016) 5:3*
DOI 10.1186/s13388-016-0028-1

Security Informatics

CASE STUDY Open Access

Analytics for characterising and measuring the naturalness of online personae

Jason R. C. Nurse*, Arnau Erola, Thomas Gibson-Robinson, Michael Goldsmith and Sadie Creese

Abstract

Introduction: Currently 40 % of the world's population, around 3 billion users, are online using cyberspace for everything from work to pleasure. While there are numerous benefits accompanying this medium, the Internet is not without its perils. In this case study article, we focus specifically on the challenge of fake (or unnatural) online identities, such as those used to defraud people and organisations, with the aim of exploring an approach to detect them.

Case description: In particular, through our method and case study we outline and experiment with novel analytics for characterising and measuring the naturalness of an online persona or identity; this naturalness is defined as the extent to which that persona has features similar to those expected for comparable personae online. Our case scenario involves a participant set of two types of individuals, and our aim at this stage is to use our approach to correctly characterise, and then distinguish between, these two types.

Discussion and evaluation: To briefly *précis* our case study results, we found that our method to conceptualise an individual's complete online presence was very successful. This was undoubtedly linked to its detailed consideration of how cyberspace is typically used, while also building on our existing model of identity which has been used to aid law enforcement in identification tasks. In terms of developing effective analytics for naturalness however, improvements in our approach (e.g., features selected and nuanced metrics) are required. Moreover, the study would benefit from a larger sample size to better identify common aspects between natural personae.

Conclusions: Overall, the case study allowed us to explore a novel technique to characterise naturalness and to examine its utility at detecting unnatural personae. Our goal now is to build on the study's findings in several key ways. Specifically, we aim to conduct further assessments on the criteria through which naturalness is defined, and refine our analytics and combinatorics to measure a persona's naturalness. We will also explore clustering approaches based on complete online personae, as a means to complement our identification of naturally occurring personae types in large datasets.

Keywords: Identity security, Identity theft and fraud, Detection approaches, Data analytics and metrics

Introduction and related work

Today more than ever, people across the world are exploiting the Internet for work and pleasure, and are utilising an increasing variety of devices and services to do so [1]. Exploitation of cyberspace results in both conscious information-sharing and publication (of both personal and corporate variety) and, inevitably, the creation of persistent data that many users may be unaware of (perhaps as metadata or as old data thought to be removed or put out of reach). While there are undoubtedly many benefits to our interaction in cyberspace, the quantity of threats, risks and general peril are constantly growing [2], with data breaches, hacks and identity-fraud almost commonplace.

In this case study-based article, we concentrate on the problem of fake online identities, and their increasing use to manipulate, deceive and defraud people and

*Correspondence: jason.nurse@cs.ox.ac.uk
Department of Computer Science, University of Oxford, Oxford, UK

organisations [3–5]. Reflecting on the literature, there has been considerable work in the space of detecting fake accounts and bots. Cao et al., for instance, propose a technique using social-graph properties to rank users on a site according to their likelihood of being fake [6]. They later extend this consideration to explore the use of clustering in identifying groups of malicious accounts (under the assumption that they have very similar properties and actions, e.g., posting and uploading behaviour) [7].

In Viswanath et al. [8], an unsupervised machine-learning approach for detecting anomalous user behaviour in social networks is introduced. Through experimentation on Facebook profiles, the authors demonstrate the use of their clustering technique (based mainly on 'like' rates and activities) in identifying fake and compromised user accounts. In the spam and bot-detection domain, Fong et al. [9] and many others (e.g., [10, 11]) also attempt to tackle the problem of fake profiles, typically using a mixture of techniques, which often apply a priori knowledge (e.g., bots ephemeral nature or posting habits) to detect fake accounts. We have also engaged in research in this domain by using machine learning to explore which factors may be the most important in making automated text (produced by bots) convincing [12].

The novelty of our work as compared to the existing literature is the in-depth analysis of a complete online persona; this includes all of its facets and how it is used across multiple sites. We posit that through a detailed characterisation of how real personae portray themselves and act online, an approach can be crafted to detect fake or anomalous identities, particularly, those somewhat carefully maintained and used for malevolent purposes. As a basis for this approach, we draw on a comprehensive model of identity developed in our previous research [13, 14]. This allows us to characterise identities, from the attributes present and the inferences that can be made from them (e.g., inferring a person's name from their email-address), to the overall existence of an identity across several sites. The ability to comprehensively model an identity can be an extremely effective tool in understanding what is natural behaviour online, and consequently, what may be an unnatural and potentially harmful persona.

In what follows, we present our approach and the case study used to examine it, before then critically reflecting on our findings regarding the approach's utility. Specifically, we first consider naturalness as a concept, what it means for a persona to be natural and how naturalness may be usefully characterised. Next we detail the analytics (i.e., intervention) that we propose for measuring the naturalness of an unknown online persona. We then present the results from a case study experiment conducted to explore our approach. Finally, we reflect on these results and outline ways to evolve the analytics, before concluding the article.

Approach and case study

We begin our work in this section by introducing the proposed approach to characterise and measure naturalness. This is then followed by a definition of our case study and presentation of results.

Defining and characterising naturalness

Online identities and naturalness

To properly consider an online identity, there are several important concepts that first need to be understood. One of the most central of these is that of a *persona*. We define a *persona* as the way in which an individual (consciously or unconsciously) presents themself online. A key defining characteristic of a persona is that it presents a largely consistent view of an individual. The way that individual actually portrays or manifests their persona online is through what we refer to as a *profile*. Profiles are typically local to a website (e.g., Facebook, eBay) and represent user accounts held by or about the individual. Profiles maintain sets of identity attributes (e.g., name, username, photos, details), hereafter *elements*, about a persona; and we also consider *inferences* that define techniques by which new elements can be derived from existing ones (e.g., based on language-analysis tools, one can assess a Facebook post and infer a person's mood or sentiment).

Another concept crucial to our discussion is that of *contexts*. The term *context* is used to represent a particular type of online space, for instance a work-related space or a space focused on socialising. Contexts are intended to provide a very broad way to characterise a set of related elements, and could also be used to conceptually describe or group such elements within a profile. At a finer granularity to contexts are *topics* or *topic areas*—these function in the same way as contexts but are more fine-grained in the related identity elements they group together. For instance, one might have a work context, and within that context have topics about projects engaged in at work or events with work colleagues; these topics would then bring together related model elements such as project descriptions and collaborators, and event location, time and attendees. Similar topics might also exist under multiple contexts.

Using the concepts above and drawing on our existing identity model [13, 14], the naturalness of a specific persona is defined as *the extent to which that persona has features similar to those expected or standard for comparable personae online; here, expected or standard features refers to the range of profiles, contexts and model elements that are present, and inferences that are achievable, in the majority of similar personae*.

To take an example, assume that we have assessed a set of related personae and from the data gathered, we have inferred what is natural for personae of this kind. Now, further assume that this naturalness is characterised by an online presence in a university web page and a LinkedIn profile; both with data on the research projects the individual is involved in, articles published, and teaching responsibilities. Therefore, if we identify a persona claiming to be of this kind (e.g., via a specific LinkedIn profile) that is not similar to our characterisation of normal above (whether it be in presence or absence of topics, availability of elements, or ability to conduct inferences), then we may assign this persona a low naturalness score; the general idea being that the lower a naturalness score, the more likely that the online presence under investigation might be fabricated.

To consider naturalness thoroughly, it is necessary to be aware of the range of profiles, contexts and topic areas in which data on a persona may be found online. In order to discover these aspects, we have engaged in an in-depth study of online identity data, while also reflecting on our existing research [13, 14]. Figure 1 presents the conceptualisation of an individual's online presence resulting from our analysis.

In detail, an individual can represent themself through online personae, which, in turn, are manifest through a wide range of profiles. These profiles can link together identity elements on practically any subject and for any purpose; the figure presents a small subset of arguably the most popular of these profiles. As a way of describing profiles and the data held within them, we hypothesise that there are at least five high-level contexts to which profiles could be associated.

These contexts are: the *social context*—identity information generated from the use of the Internet as a social medium (e.g., persona Facebook profiles, personal life blogs, gaming); *work context*—identity information as a result of Internet usage for official work and employment purposes (e.g., LinkedIn profiles or a company's employee page); *state context*—this covers identity data on a country's citizens typically provided by a government as a result of online initiatives (e.g., Electoral rolls, Civil registries); *customer context*—identity information that arises due to the use of the Internet for purchasing and providing reviews on goods and services (e.g., Amazon or eBay profiles); and *community citizen context*—identity information pertaining to the use of the Internet for participating in a support community (e.g., neighbourhood watch, volunteering). Although we present contexts under profiles, they also can be presented above, or indeed as a profile annotation. A profile can be associated with multiple contexts or one context can describe data in multiple profiles.

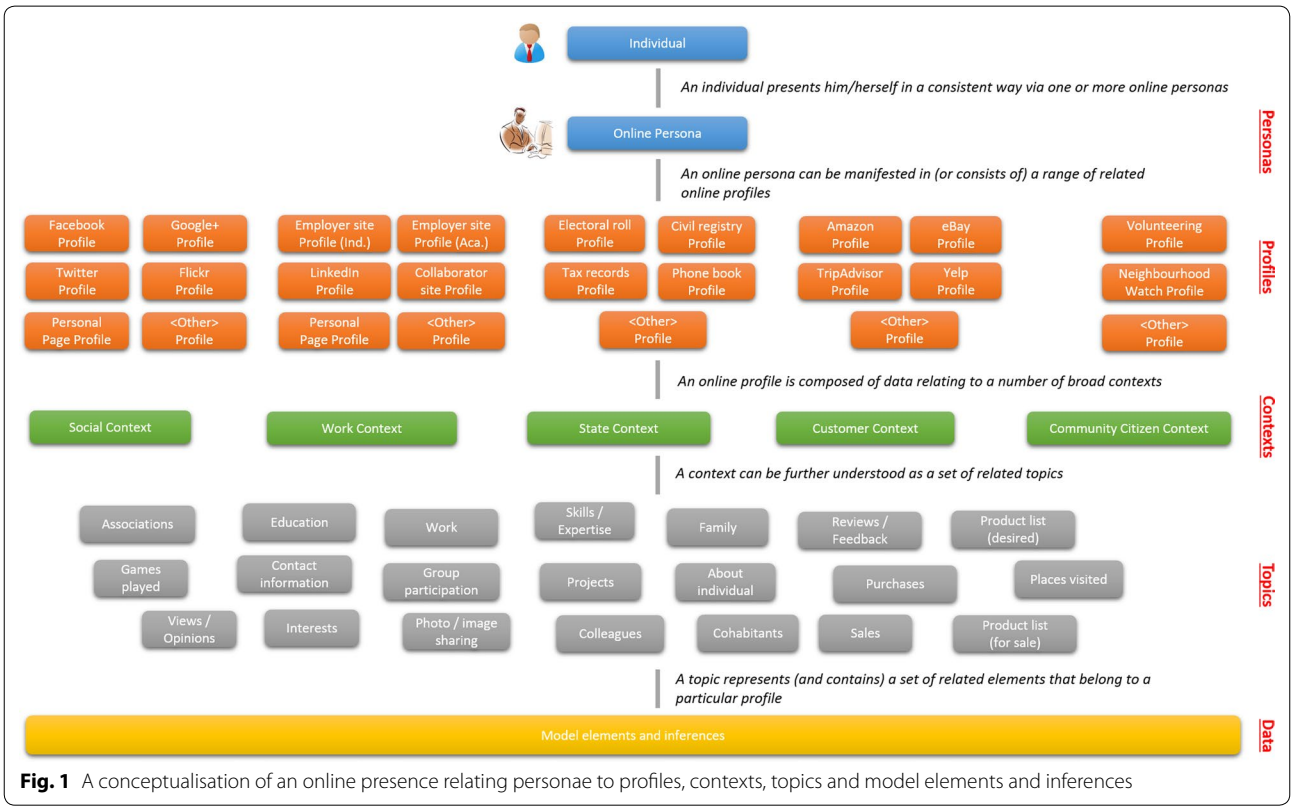

Fig. 1 A conceptualisation of an online presence relating personae to profiles, contexts, topics and model elements and inferences

Under the contexts layer in the diagram, there is the notion of topics (or, topic areas). The 'Interests' topic for instance, aims to capture behaviour in a profile that pertains to an individual's interests, likes, dislikes, etc. If we apply this to a Facebook Profile, this topic would encompass identity elements such as *Likes*, *Movies*, *Music* and *Books*, and identity-model inferences covering what could be derived from those elements; for instance, interests in specific books might lead to insight into an individual's expertise or family life. This highlights one of the advantages of the proposed conceptualisation: a profile can contain any of the listed contexts and topics, while topics can draw on, and be constituted by, a variety of elements and inferences at the lowest model layer. In Fig. 2, we provide an example of the detailed mapping of the elements and inferences from our identity model, to the related topic areas.

Using the conceptualisation for naturalness

There are two ways in which we could look to apply the conceptualisation embodied in Fig. 1 to characterise naturalness. The first method is to adopt a top-down or static perspective to analysis, and thus to be guided largely by the existing structure in that diagram in the search for and classification of an identified individual's data. Figure 3 depicts an example of how the conceptualisation can model the online presence of an individual.

To define naturalness for a set of individuals or personae of a specified type therefore, this approach would be guided mainly by what personae, profiles, contexts, topics and elements tend to be available, and the inferences that are usually possible. Therefore we are interested in questions such as: whether certain profiles (e.g., Facebook, Twitter or 192.com) are commonly available, whether existing profiles are typically used for particular reasons (i.e., fit certain contexts or topics), and ultimately, whether there are any elements that are usually shared by the individuals or inferences from one element to another that can be expected. These will be the crucial factors in understanding and characterising naturalness especially in the context of our identity model.

The second method of analysis is to apply a bottom-up and more dynamic perspective to the problem. The aim here is to start at the bottom of the layout with the data elements of a supposed individual, and then to group these elements by how they are related, and continue to build upwards. The first step, therefore, would be to use the data elements (from the person's profiles) to construct actual topics present for that specific individual. Once the topics have been identified, we would then move to group related topics into high-level contexts– these contexts would provide general insight into how the profile is actually being used by the individual. To move upwards from profiles to personae (i.e., to determine whether profiles belong to one persona or many personae), this approach proposes to rely on a set of core elements of a profile and group profiles into the same persona if these elements are jointly similar. These core elements could include: name, age, location and email-address. At this point, this technique has been validated on a small dataset with five individuals.

Similar to the top-down approach, naturalness in this more dynamic approach would be characterised by considering what elements, inferences, topics, contexts and profiles tend to arise (and how or when they tend to arise) across a set of individuals and their personae. The advantage with this approach is that it allows for the discovery of new topics, contexts and profiles not previously thought of.

Topic area	Related elements	Related inferences
About individual	age, birthdate, biography, gender, language, name, links to other social-media accounts, username,	age ⇒ birthdate biography ⇒ interest links to other accounts ⇒ website profiles name ⇒ ethnic origin, gender, username, website profiles username ⇒ age, email-address (personal), gender, name, personality traits, website profiles
Interests	likes, favourites, user/content tags	favourites ⇒ age, gender, personality traits, personal views likes ⇒ age, employer, expertise, gender, personality traits, personal views
Work	current employer(s), current job title(s), job time periods, occupation, past employer(s), past job title(s),	current/past employer(s) ⇒ address (work), email-address (work), employer website, individual's location, phone number (work), username (work) current/past job title(s) ⇒ expertise, occupation
...	...	...

Fig. 2 Low-level data mapping An excerpt of the mapping of identity model elements and inferences to topic areas

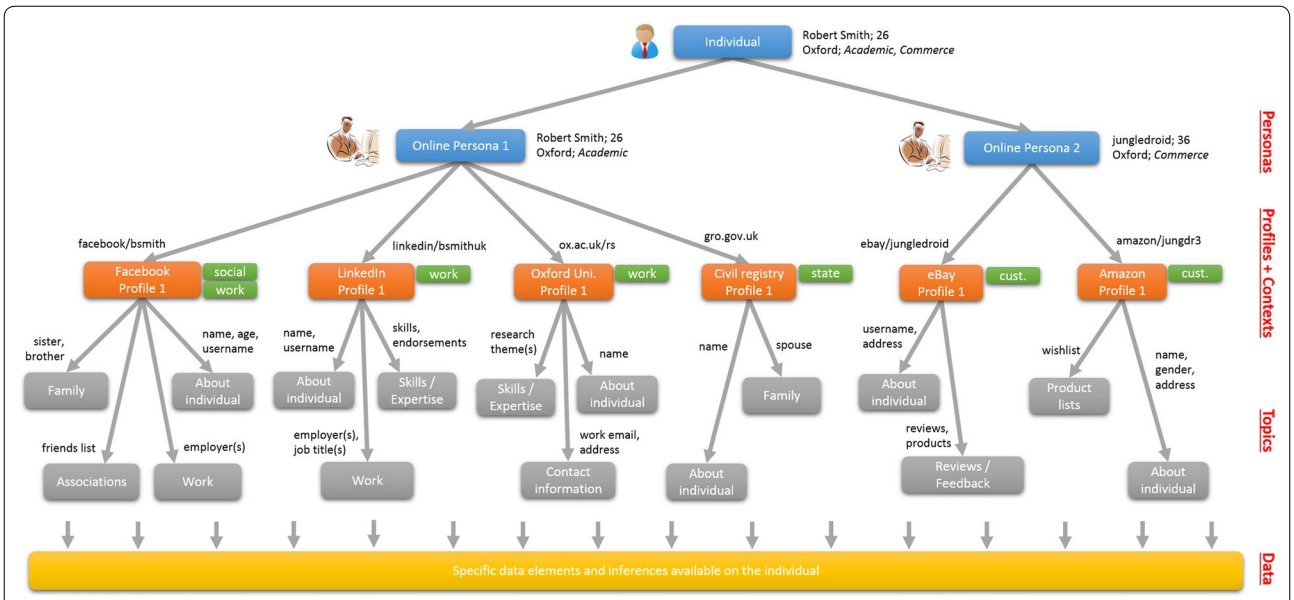

Fig. 3 Applying the conceptualisation An example of how we might capture and model data about an individual. These personae portray the individual as an academic (*Robert Smith, 26, Oxford*) and as someone engaged in online commerce (*jungledroid, 36, Oxford*). The personae are manifested through several online profiles on sites such as Facebook, LinkedIn, eBay, and a Civil registry

Analytics for measuring the naturalness of a persona

Our approach to measuring a persona's naturalness is composed of two main steps, characterisation and assessment of naturalness. These are detailed below.

Research and characterise what constitutes naturalness for a set of individuals and their personae

This task involves several smaller naturalness-characterisation activities:

Step 1 Identify a set of individuals and collect the identity information available on each of them in the online space. Practically, this would start at the profile level, and would therefore involve noting the profiles maintained by the individuals. In terms of obtaining the most value from this characterisation task, it is useful to choose individuals of a particular type (or sets of types) where there may be some plausible commonality in their online presence and personae. In addition, this choice may be guided by the types of (potentially) fake persona that may be investigated later in the assessment.

Step 2 For each individual, apply the conceptualisation from the previous section (in Fig. 1) to create models of their online presence. This task could use either the top-down or bottom-up approach, but will need to be consistent across the individuals. The structures resulting from this analysis would identify the profiles, contexts, topic areas and elements present, along with the inferences possible, for each of the individuals' personae; the mapping in Fig. 2 would be useful here, especially in the top-down analysis. This step will also note the *values* of certain elements, such as posts and images, in addition to summary values such as number of friends, favourited items and posts per day; these will be used in later characterisation tasks. As an output from this step, we would expect several layouts similar to that in Fig. 3.

Step 3 Analyse the set of structures emerging from individuals' data to determine whether there are certain aspects that commonly or naturally arise in the structures; these would then be documented. Key questions would be: are certain profiles, contexts or elements mostly present, or never present? Also, can certain inferences commonly be performed, or are they largely impossible? The identification of inferences expected to be possible is a crucial step, as we expect that unlike many of the other steps above, great variance (at least, across similar individuals) is unlikely. Another important task here is assess the data elements available on individuals to determine whether it is possible to define any normally occurring values for the elements themselves. Again, this would look to use what happens in the majority of cases. To take Twitter as an example, this task would look to identify whether or not it is natural for individuals to select an image of themselves, or others for their avatar. We suspect that there are a few control factors that may need to be taken into account during this step's analyses. For instance, the age of an individual might have notable influence on how their personae manifest across the spaces mentioned.

Step 4 Summarise the findings of the earlier steps (2–3) and create a template of naturalness for that set of individuals and their personae. This could adopt a layout similar to that in Fig. 3 but instead, the presence of elements, topics, contexts and profiles, would be dictated by whether they could be expected to be present. As with several of the steps above, this characterisation of natural will be based on what occurs within a majority of the individuals, but not necessarily all of them. With this in mind, we might need to be slightly flexible in our definition of (or tolerance for identifying) naturalness, especially in situations where we were unable to find a clear majority behaviour.

Assess the naturalness of the persona that is under investigation

In detail:

Step 1 Compare and contrast the persona of interest against the naturalness template of personae previously characterised. This step assesses the extent to which the new persona's online presence (structured as in Fig. 3) is similar to the expected presence of comparable personae. A crucial factor here is that the persona to be assessed is largely of the same type as the personae earlier used to characterise naturalness. This would ensure that the measurement approach distinguishes unnatural personae and not just ones that are of a different type. We propose to assess the naturalness of a new persona based on a measure of overlap in expected profiles, contexts, model elements that are present (or absent as the case may be) and inferences that are not enabled. To conduct this measurement, we have explored several approaches, of which two are outlined below.

The first approach favours simplicity and directly compares the defined conceptualisation of the new persona to a natural persona template. This comparison is conducted on a per item basis with each element, context, profile, and inference being compared. If the two items agree, we assign 1, otherwise 0 is assigned. For instance, if it is natural for individuals to state their school history on LinkedIn, as well as posting an image of themselves as their avatar, yet the persona under investigation only does one of these, they would receive [0, 1]. These values are then averaged across the full set of items to get the persona's percentage similarity to the naturalness template. If that average is less than some predefined naturalness threshold (which could be adapted based on the sensitivity of the assessment), then the persona would be deemed unnatural.

The second approach to measuring naturalness is to factor in the percentage of agreement with the naturalness template. This would consider the fact that although the template represents the majority value, it could well be ignoring some level of disagreement that is important in subsequent measurement tasks. The approach is as follows. Firstly, define the structure for the new persona similarly to Fig. 3. Next, compare each item in the persona with the respective item in the naturalness template. A penalty, or disagreement to the template, is then calculated. In cases were there is agreement, assign 0 thus no penalty, otherwise, a penalty between 0 and 1 is to be calculated for that item. Penalties are defined based on the extent of disagreement to the norm in the initial set of individuals.

There are a number of ways in which penalties might be derived. One way is to use a linear approach where the penalty for the new persona not agreeing with the natural value is directly proportional to the percentage of agreement with the natural value (in the initial set of individuals). For instance, if in characterising naturalness it was found that only 51 % had agreed with the selected natural value, then if a persona under investigation also does not agree, they should be penalised considerably less than the case where 98 % of people agreed. To achieve this, we set penalties to increase on a linear scale from 0.02 (where there is 51 % agreement) up to a penalty of 1 (at 100 % agreement).

In addition to the linear approach, we have also explored the application of logarithmic, exponential and power functions to produce penalty values. These use a broadly similar method as above to incorporate the percentage agreement but, as is to be expected, output different sets of penalties. Once values (and penalties) have been applied to each item to indicate similarity to the template, these are averaged to determine the similarity of the persona itself to the natural persona. Depending on the threshold set, a conclusion as to the potential naturalness of the persona can be reached.

Case study experimentation and results

To assess the ability of our approach to characterise and measure naturalness, we conducted a case study experiment. At this stage, we felt that a case study would be more appropriate, as were especially interested in (a) exploring the characterisation process and (b) in understanding whether the approach could at least detect personae (and thus, individuals) of the same or differing types. We view these tasks as key prerequisites in being able to detect fake personae. As such, we designed a case scenario with a participant set of two general types (students to characterise natural, and professionals to test it). Here we did not attempt to detect genuinely fake personae and did not include an assessment of known fake personae—this would form the basis for our next and more comprehensive experiment. Below, we present the case study experiment and then highlight some of the main findings.

Characterising naturalness: Step 1

To characterise naturalness, we started by recruiting 30 students as a basis for our case study. We hypothesised that given they were of the same age and studied very similar degrees at the same institution, they *may* have a generally similar online presence. With the permission of the participants and in line with university ethical guidelines, we then collected their online identity data, including profiles, identity elements, source data, and so on.

Characterising naturalness: Step 2

Next, we analysed the online presence of each individual and created several models similar to Fig. 3. While the definition of profile, topic areas, elements and inferences was straightforward, specifying appropriate contexts required a comprehensive analysis of the profile elements of each individual and determining how that profile was being used, e.g., for work, social, and so on. In general however, this step progressed as expected with the conceptualisation more than capable of adequately abstracting each presence.

Characterising naturalness: Step 3

The various identity conceptualisations were then assessed with the aim of ascertaining what profiles, contexts, topic areas, elements and inferences may be regarded as natural. Below, we summarise the analysis.

Defining natural profiles To define naturalness at the profile level, we adopted a simple approach based on what occurred in the majority of cases. That is, if most personae possessed the profile, this was considered as natural for this sample, and if most did not, that was considered as natural. Figure 4 presents a summary of the different profiles maintained and their prevalence.

One notable finding was that only 6 of 31 profile types could be considered as naturally present for the participants, i.e., at least 50 % of the participants had them. These were Facebook, Twitter, Google+, YouTube, Amazon and eBay. Some other profiles did have a reasonably strong presence (e.g., Instagram, the Steampowered games platform, Spotify and LinkedIn) but not a majority. In contrast, there are some profiles which are very uncommon, so the natural tendency is not to possess them (e.g., IMDB, Stackoverflow).

Another point worth highlighting is variability in the total number of profiles maintained by personae. Overall, the average number of profiles is 8, with a standard deviation of 4. A question that might arise here, therefore, is whether or not certain individuals within our initial naturalness characterisation set were outliers.

Defining natural contexts The definition of natural contexts utilised the majority approach as applied above. Reflecting on the analysis, the contexts found were somewhat expected given participants' background. For instance, Facebook, Twitter and Instagram all were used to socialise (i.e., interacting with friends, posting pictures, and so on), rather than for work or business, and therefore these were naturally associated with the social context.

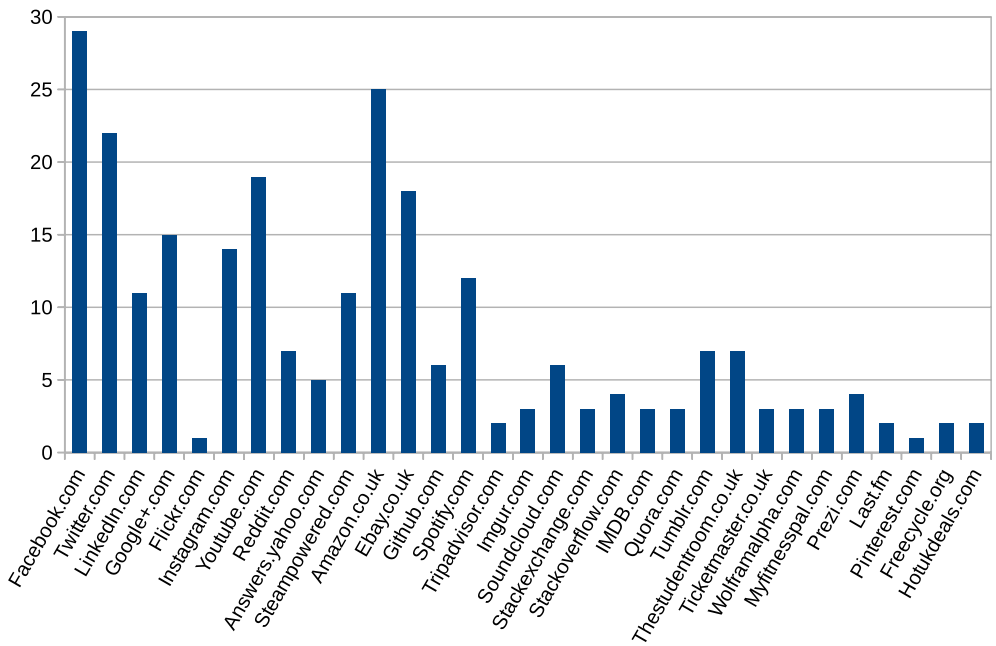

Fig. 4 Identification of maintained profiles number of accounts that the participants of the study maintain in various sites

Unsurprisingly, eBay was naturally characterised in the customer context. An interesting point that arises here is that with some sites, they likely only have one use or arguably a main use, and therefore may not be as useful in distinguishing unnatural personae at a context level.

While *topic areas* were used in our analysis, we do not report on them here as their findings were very similar to the element-level assessment (next).

Defining natural elements This step sought to determine which identity elements were naturally (mostly) exposed in profiles. As before, we characterised the naturalness template for each profile's elements, but for brevity here we only present results for Facebook. In Fig. 5, the availability of elements across individuals can be observed. A key finding here was that there are some elements that were always present because they are required (e.g., *gender*), and others that are just mainly used by individuals (e.g., *avatar*, *coverpicture*); it is the latter of these which we use to define naturalness. There are also some elements that have a clear lack of prevalence (e.g., *website*) and therefore it could be concluded that it is natural for personae of this type not to link to other profiles.

Defining natural inferences Similarly to the process with identity elements, we characterised inferences that were naturally possible and impossible. This process progressed as expected and we were able to define natural inferences, but it required a notable amount of manual effort in assessing the ability to conduct inferences. For example, on Facebook, we discovered that participants have a clear tendency to use a picture of themselves as their avatar (97 %), sometimes appearing with others (62%). We also found that the posts on participants' walls naturally mentioned friends, locations, liked organisations or companies, and the individual's personal interests.

Defining natural values Another approach adopted to ascertain naturalness was to analyse the *values* of elements. We focused on two areas: (a) textual content analysis to assess aspects such as whether or not it is natural to maintain consistent values (e.g., *age* or *name*) across profiles or if values typically tend to match the ground truth (e.g., real age); and (b) frequency analysis of numeric elements (e.g., number of *posts*, *pictures uploaded*) to calculate the range of potentially natural values (assuming a Gaussian distribution, 'natural' ranges could be the average value plus/minus the standard deviation). Findings from (a) included the fact that there is a clear majority tendency for individuals to have the same age and gender across profiles, but not similar locations or usernames. As it pertains to (b) and Facebook for instance, we found that personae tend to post more during the evening and less during the morning, but there is not a specific hour where they mostly publish. Therefore, if we were to find an individual who always publishes at the same time, this could be regarded as unnatural.

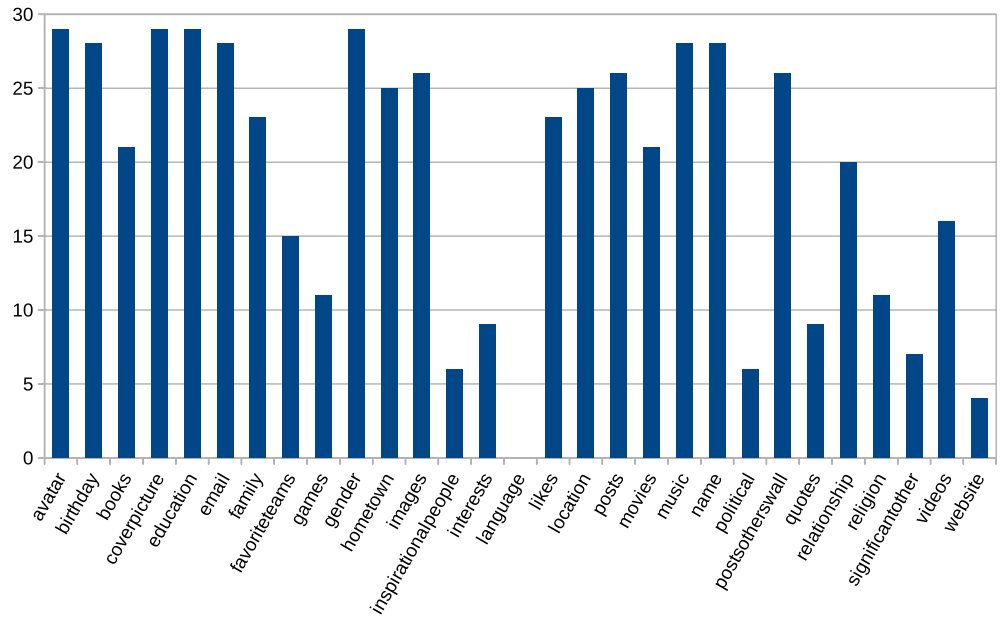

Fig. 5 Identification of model elements in Facebook Number of model elements that are filled for the Facebook accounts of the participants of the study

Characterising naturalness: Step 4

From the full analysis of participants' data, we defined the high-level natural persona template in Fig. 6. This highlights the profiles, contexts, data elements and inferences that could be expected to be available for a natural persona of the type under investigation.

Measuring naturalness of new persona

This section applies our general approach to measure the naturalness of new personae; in the context of our wider research aims, these would be the suspected fake personae. For our case study analysis, we used five new personae, three of which were students of the same degrees as the initial set and therefore may be expected to be found as natural when compared. The other two personae were from older individuals and employees of companies in different subject areas; as such their presences might appear unnatural or at least different to the students' natural personae template. Hereafter, Test personae #1, #2 and #3 are the students and #4 and #5 are the professionals.

Measuring naturalness at the profiles level We began by evaluating how the five test personae compared to our naturalness template in terms of profiles. Figure 7 displays an example of this evaluation for Test persona #1. For each profile type, we have documented its natural behaviour (i.e., whether it is natural for individuals to possess it [1] or not [0]) and the percentage of agreement to that norm by participants within the initial set (where 100 % indicates complete agreement and 51 % defines very limited agreement). Next, we took the list of sites where the Test persona #1 had a presence, and the calculated naturalness metrics. Specifically, we used a simple penalty metric (where 1 is for agreement to the natural template and 0 otherwise), a penalty metric using a linear curve, and a penalty metric using a power curve. As can be seen in the figure, generally the simple penalty obtained the greatest values as it equally penalised all the differences with the template. On the contrary, the linear and power approaches penalised according to the percentage agreement of the initial participant set.

For the specific persona in Fig. 7, we can observe that it is very similar to the natural template (i.e., over 93 % similarity across all metrics); with similarity defined as [*1-AverageMetricScore*] where *AverageMetricScore* is the average of the penalties across all profiles.

To comment on the general results at the profiles level, we were unable to use our approach to distinguish between students and employees (thus, natural and unnatural personae). This could be because the type of personae represented by individuals #4 and #5 is very similar to our natural profile template, or because there was too much variation in our sample used to define naturalness. To test this last hypothesis we adopted the approach mentioned in *Step 1* that focused on defining and reflecting on the average penalty and the standard deviation for the initial participant sample. From this

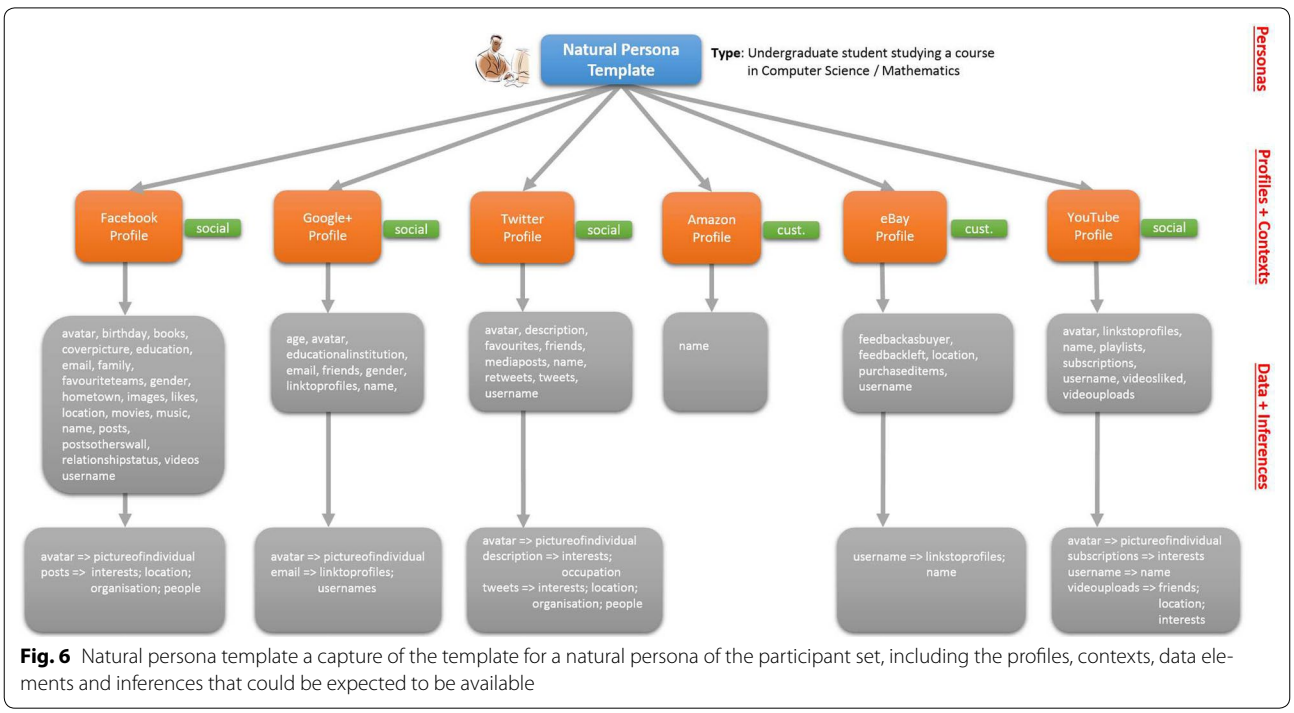

Fig. 6 Natural persona template a capture of the template for a natural persona of the participant set, including the profiles, contexts, data elements and inferences that could be expected to be available

Profiles:	Natural template	% agreement with Nat. template	Test Persona #1	Metric 1 (Simple)	Metric 2 (Penalty - linear)	Metric 2 (Penalty -power)
Facebook.com	*1*	*97%*	1	1	0	0
Twitter.com	*1*	*73%*	1	1	0	0
LinkedIn.com	*0*	*63%*	1	0	0.26	0.000216
Google+.com	*1*	*50%*	1	1	0	0
Flickr.com	*0*	*97%*	0	1	0	0
Instagram.com	*0*	*53%*	1	0	0.06	0.017576
Youtube.com	*1*	*63%*	1	1	0	0
Reddit.com	*0*	*77%*	0	1	0	0
Answers.yahoo.com	*0*	*83%*	0	1	0	0
Steampowered.com	*0*	*63%*	0	1	0	0
Amazon.co.uk	*1*	*83%*	1	1	0	0
Ebay.co.uk	*1*	*60%*	1	1	0	0
Github.com	*0*	*80%*	0	1	0	0
Spotify.com	*0*	*60%*	0	1	0	0
Tripadvisor.com	*0*	*93%*	0	1	0	0
Imgur.com	*0*	*90%*	0	1	0	0
Soundcloud.com	*0*	*80%*	0	1	0	0
Stackexchange.com	*0*	*90%*	0	1	0	0
Stackoverflow.com	*0*	*87%*	0	1	0	0
IMDB.com	*0*	*90%*	0	1	0	0
Quora.com	*0*	*90%*	0	1	0	0
Tumblr.com	*0*	*77%*	0	1	0	0
Thestudentroom.co.uk	*0*	*77%*	0	1	0	0
Ticketmaster.co.uk	*0*	*90%*	0	1	0	0
Wolframalpha.com	*0*	*90%*	0	1	0	0
Myfitnesspal.com	*0*	*90%*	0	1	0	0
Prezi.com	*0*	*87%*	0	1	0	0
Last.fm	*0*	*93%*	0	1	0	0
Pinterest.com	*0*	*97%*	0	1	0	0
Freecycle.org	*0*	*93%*	0	1	0	0
Hotukdeals.com	*0*	*93%*	0	1	0	0
Similarity of new persona:				93.55%	98.97%	99.94%

Fig. 7 Similarity with the natural template similarity metrics of a persona with the calculated natural template for having accounts in different sites

analysis, we found that at this level, there were six participants that had a penalty score higher than the mean penalty (0.0343) plus one standard deviation (0.0334) and one participant (of that six) with a score higher than the mean plus two standard deviations (unnatural). In some ways, these individuals could be considered as outliers whose profiles could have weakened the naturalness characterisation process.

Measuring naturalness within profiles Here we assess the naturalness of test personae in terms of contexts, profile elements and inferences. Due to limited space, we focus on Facebook as a findings example.

At the contexts level, our approach was able to detect differences in Test persona #4 and #5 as compared to the natural template. This was because these personae had several details about their work (e.g., *profession*, *employers*) on their profiles. In terms of the overall similarity score at this level however, the metrics still deemed these two personae as natural.

To measure the naturalness of test personae at an elements level, we drew on the characterisation (from Fig. 5), and followed the methodology defined in Step 3. We started by calculating the threshold and standard deviation for the similarity of the initial set, which were 74.51 and 7.89 % respectively. Thus, we considered individuals to be unnatural when their similarity (to the template) is lower than 74.51 %. Using this threshold, all test five personae appear natural. This could again be the result of the high variability in the initial dataset thus affecting the deduced natural template. Alternatively, it might be the result of considering too many elements, and thus, reducing the mean score and introducing noise. The thresholds for the linear and power metrics are 89.26 and 95.19 % respectively. Only Test persona #1 has average similarities lower than these, i.e. indicating that the data within their profile is potentially unnatural. This is the effect of the penalties for *hometown* and *location* elements in particular, which this individual has not provided.

The next task is to consider the naturalness of inferences in Facebook. In Fig. 8 we can see a survey of the thresholds for Facebook and the similarities that three of the test users receive. The calculated thresholds are 61 % for the simple penalty, 83 % for the linear penalty and 94% for the power penalty. According to these metrics, Test persona #2 looks unnatural using the simple and linear penalties because they are not exposing *locations*, *organisations* or *interests* in their posts. Similarly, Test persona #4 looks unnatural particularly as they do not expose *organisations* and *interests*.

To briefly consider naturalness at an element-values level, we compared findings from the new personae with the textual content and frequency analyses. For the textual analysis, the professional test personae were generally different to the natural template, especially in username consistency across profiles. For the frequency analysis, Fig. 9 presents the time of the day the test personae tended to publish on Facebook. The dark shadow represents the mean plus one standard deviation (i.e., naturalness threshold), while the light shadows is the mean plus two times the standard deviation. Here, Test persona #1 (blue) appears unnatural as they publish very often early on mornings, while Test persona #4 (green) looks unnatural as that individual has a tendency to publish most around 2pm.

Next we reflect on our case study analysis and the ability of the proposed approach to achieve its aims.

Reflecting on the approach and analytics

In this research, our aim was to develop a method that could characterise the natural online presence of a type of individual and analytics to measure whether personae of unknown origin might be considered as natural. In general, from our case study analysis, we found that our approach to conceptualise an individual's presence could be regarded as successful. In terms of developing effective analytics for naturalness however, improvements in the approach are required. Ultimately, this meant that we were not able to use the measure as is it stands to distinguish natural from unnatural (or differently-typed) personae. Consequently, this has affected the planned subsequent use of the approach to detect fake online persona. Below, we reflect on some of the main reasons why this might not have been possible.

The dataset

In any approach to identify fake or unnatural personae, there must first be some clear understanding of what is natural. We believe that natural behaviours can change across different types of individuals and thus, identifying the type of individual or personae to be assessed is particularly important. We approached this issue by defining and exploring a case study, which involved recruiting a set of students of the same age and studying for the same degree. Unfortunately, in our analysis of their online

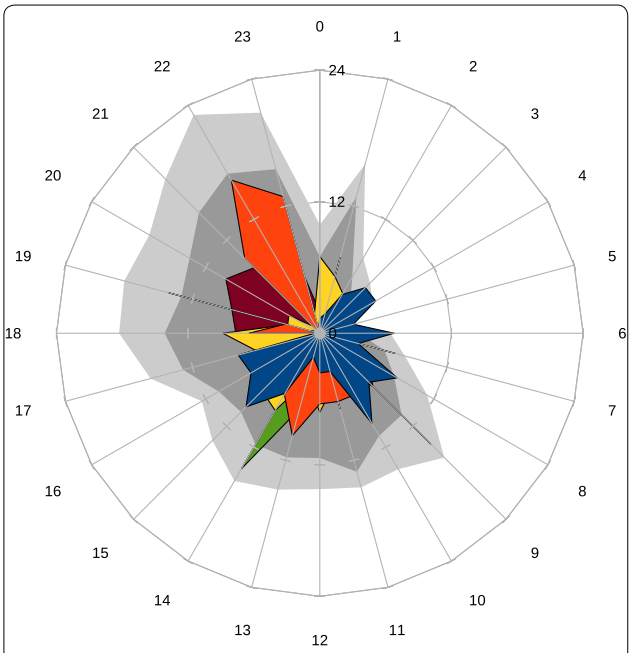

Fig. 9 Frequency of posts in Facebook for personae a summary of the frequency of posts in Facebook during the day for the five test personae (*coloured*). The *dark shadow* represents the mean plus one standard deviation (i.e., naturalness threshold), while the *light shadows* is the mean plus two times the standard deviation

	Naturalness threshold			Test Persona #1			Test Persona #2			Test Persona #4		
	Metric 1 (Simple)	Metric 2 (Penalty - linear)	Metric 2 (Penalty -power)	Metric 1 (Simple)	Metric 2 (Penalty - linear)	Metric 2 (Penalty -power)	Metric 1 (Simple)	Metric 2 (Penalty - linear)	Metric 2 (Penalty -power)	Metric 1 (Simple)	Metric 2 (Penalty - linear)	Metric 2 (Penalty -power)
Profiles	67%	91%	91%	94%	99%	100%	68%	83%	91%	90%	98%	100%
Elements	75%	89%	95%	79%	88%	94%	83%	94%	99%	83%	94%	99%
Inferences	61%	83%	94%	78%	93%	98%	56%	78%	93%	78%	86%	95%

Fig. 8 Survey of similarities similarities obtained using the three different metrics for all the test personae according to profile, model elements and inferences in Facebook

data we found notable variations in their presence which undoubtedly impacted the naturalness persona template deduced. One conclusion from this is that types based on profession or age may not be best for defining naturalness.

The sample size could have also had an impact on our analysis and findings. Due to the small size of the case scenario dataset, the calculation of true penalties may have been incorrect or focused on elements which were not relevant for naturalness (or vice versa, i.e., omitting elements which are important). Overall the method seems to behave correctly if there is a clear majority in the data, so we might assume that if we had a larger sample, the method would be able to produce a more accurate template of naturalness. This is definitely an area for future work when we expand from this case study-based experiment to a substantial, large-scale study.

Naturalness characterisation

To characterise naturalness, our analytics approach has made the assumption that the majority case is the natural case. Considering our data, this meant that because a majority of individuals did not have a profile on Spotify for instance, it was not considered as natural. As a result, there was arguably no need to consider the information shared within Spotify or the associated contexts, or the inferences possible. Another way that we could approach this problem however, is to lower the threshold at the profile level, such that we would assess the data within a profile if that profile type was maintained by at least x% (e.g., 40 %) of individuals. This would allow more profiles to be included in the assessment, which could allow extra detail that might, in turn, enable an unnatural personae to be identified. We could imagine applying this approach as a secondary method if the majority-value technique does not enable unnatural personae to be discovered.

An alternative way to approach the characterisation of what is or is not natural is to only consider something as natural if it is prevalent in large majority of the initial set; for instance, in at least 80 % of cases rather than 51 %. The idea here is that some variability in the initial set should be expected and thus, we should be more strict in what is deduced as being required in order to be viewed as natural. If we take data in Fig. 6 as an example, we would only consider Facebook and Amazon as profiles that are naturally present. If we are measuring the naturalness of a new persona therefore, and they do not have a Twitter profile, they would not be penalised. This is the contrary to our current approach where the individual would be penalised quite heavily depending on the specific metric that has been applied.

A notable issue faced within the study was the variability in the data used to characterise naturalness. While this could be due to the type of personae chosen, another possibility is that there are distinct sub-clusters in this set which could better define the personae norms. To conduct a preliminary test of this theory, we applied the K-Means clustering approach [15] to the profile-level data of the initial set of personae. Our analysis found three clusters of individuals as depicted in Fig. 10; we drew on existing work to define appropriate values for K [15, 16]. While these particular clusters are of varying strengths, finding such a wide spread in the initial dataset somewhat reinforced our belief that an approach which accommodates several naturalness templates (deduced from the clusters identified in the initial dataset) may be the best way to proceed in the future. The idea here, therefore, would be to characterise naturalness via multiple templates (identified, potentially through clustering), and then to measure the naturalness of a persona of unknown provenance by assessing the extent to which it fits the templates known to be natural. If we are able to comprehensively describe the templates found, there may also be the option of identifying which templates may be best used to assess a new persona. This is a prime area for exploration in our future research.

Another general factor worth noting is that our naturalness characterisation thus far is in some regards static, and represents naturalness at a single point in time. Naturalness, however, even for the set of personae assessed, may very well change and therefore, it is crucial that the naturalness template is suitably updated. Furthermore, there is the reality that naturalness can be considered dynamically (i.e., over time), and not necessarily only at a specific point. Therefore, from repeated captures of personae data, it might be possible to identify that over a period of time, natural personae (and their respective profiles, context, and so on) tend to be characterised, or act in certain ways. Consideration of these and the other factors mentioned would allow for a more accurate assessment of naturalness in the subsequent measurements stage.

Naturalness measurement overall

Our analytics to measure the naturalness of each personae operates on a level-by-level basis, i.e., profiles, contexts, topics, elements and inferences. However, as was discussed, in order to have a general overview of the naturalness of the individual, those values should be combined in some way. Our combinatorics thus far has adopted an averaging approach, which provides a similarity score, but arguably, a rather rigid one. Another way in which overall naturalness could be defined based on these layers is as done in Fig. 6. The idea here is that we combine values at the lower level with those at the layer directly above, building a tree. For instance, when we

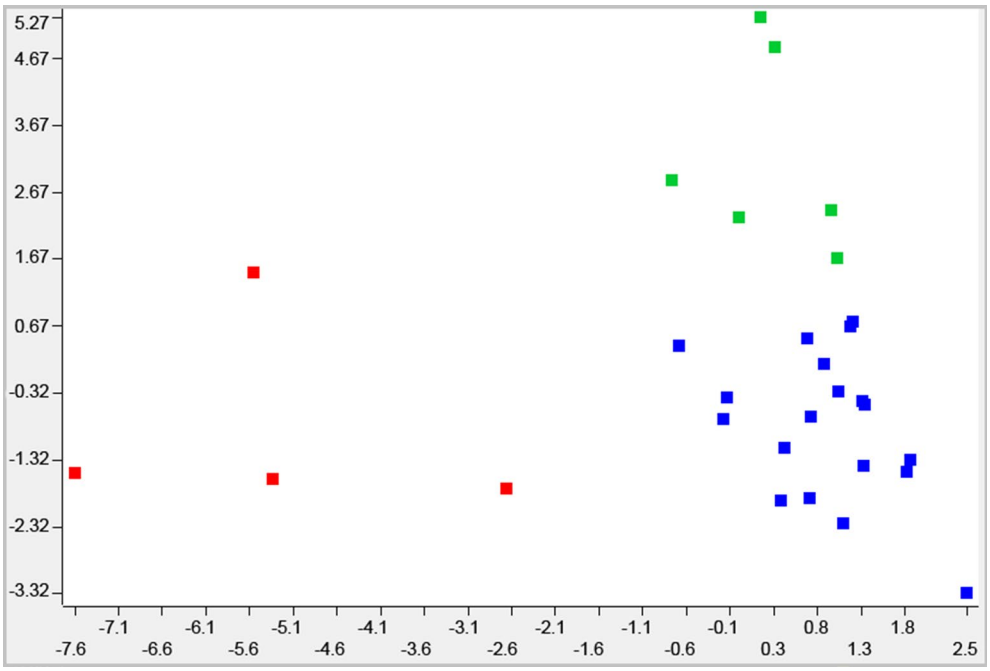

Fig. 10 Exploring the clusters of the initial set of personae a scatter plot displaying the personae of the initial dataset and clustered according to the similarity of their profile-level data. To present the vast amount of profile features on this graph, we used principal components analysis (PCA) [17] to reduce the feature space to two dimensions

calculate the average penalty for a profile's set of elements and inferences, we could then combine (e.g., multiply) it with the penalty of having that profile (if any). Below, we examine how this approach might be applied in a number of cases.

Assume that it is natural to possess a profile and the penalty for not possessing it is 0.8. Within the profile it is natural to have a specific element (*element1*) available and to make an inference (*inference1*); penalties for not conforming to the norm are 0.7 and 0.9 respective. Now, two new personae (*Persona1* and *Persona2*) are presented and we need to assess their naturalness. *Persona1* does not have the profile, and *Persona2* has the profile but *element1* is not available and *inference1* cannot be made. In the case of *Persona1*, the assessment is simple, i.e., we would just assign a penalty of 0.8 for not having the profile.

For *Persona2*, one approach is to average the penalties within the profile (i.e., $(0.7 + 0.9)/2$) and then multiply this by the penalty of not having a profile, resulting in a final penalty of 0.64— here the penalty of not having the profile essentially acts as a weight. The advantage of this approach is that if an individual possesses a profile but it is very unnatural (i.e., tends towards an average elements/inferences penalty of 1) then this would be equated as similar to not having the profile at all. In the other two cases, i.e., when a profile is not natural but the persona has it, and when the persona is not natural and the persona also does not have it, the assessments are simple. That is, in the former case the penalty for not having the profile is assigned (e.g., 0.8 in our example above) and in the latter situation, there is no penalty. Future work will need to explore this further, ideally with the larger and more clearly typed participant set.

Conclusion and future work

As the number of organisations and individuals online increases, cyberspace becomes an even more attractive area for malevolent parties, armed with various schemes and tricks meant to deceive others. In this paper, we have presented and explored an approach that is ultimately targeted at enabling us to better distinguish between real and fake (or malicious) online identities. This approach focuses on allowing an enhanced understanding of online personae, while also facilitating the characterisation of a natural online presence and the measurement of conformity to such a presence.

Reflecting on the case study-based assessment of the approach that was conducted, there were several areas where our approach performed well, but also many others where further improvement is required before it could be applied to judge fake personae. These areas will be the focus of our future work, and include: further assessments of the criteria through which naturalness

is defined, and refined analytics and combinatorics to measure a persona's naturalness. Lastly, we are in the process of exploring the full application of clustering approaches using complete online personae (i.e., data from multiple sites) as a means to identify naturally occurring personae types in large datasets. This could be used to complement our existing approach and provide more insight into the initial dataset from which naturalness (via naturalness templates for instance) would be defined.

Authors' contributions
All authors listed contributed in the research and experimentation leading to this article, and the preparation of the manuscript itself. All authors read and approved the final manuscript.

Authors' information
JN is a Cyber Security Researcher and Junior Research Fellow; AE is a Cyber Security Researcher; TGR is a Junior Research Fellow; MG is a Professor of Computer Science; SC is a Professor of Cybersecurity and Supernumerary Fellow. All authors are based in the Department of Computer Science at the University of Oxford in the UK.

Competing interests
The authors declare that they have no competing interests.

Received: 7 July 2016 Accepted: 24 August 2016
Published online: 08 September 2016

References

1. Mashable (2014) U.S. adults spend 11 hours per day with digital media. http://mashable.com/2014/03/05/american-digital-media-hours. Accessed online 25/04/2015
2. Ponemon Institute (2015) 2014: a year of mega breaches. http://www.ponemon.org/local/upload/file/2014 The Year of the Mega Breach FINAL 3.pdf. Accessed online 25/04/2015
3. Wall DS (2013) Future identities: changing identities in the UK—the next 10 years/Identity Related Crime in the UK. Technical report, UK Government's Foresight project
4. Wired (2014) How to hack governments using social media. http://www.wired.co.uk/news/archive/2014-05/29/iranian-hack-facebook-military. Accessed online 25/04/2015
5. Geek Wire (2014) How online scammers created a fake identity using little more than my picture. http://www.geekwire.com/2014/no-i-in-imposter/. Accessed online 25/04/2015
6. Cao Q, Sirivianos M, Yang X, Pregueiro T (2012) Aiding the detection of fake accounts in large scale social online services. In: Proceedings of the 9th USENIX conference on networked systems design and implementation
7. Cao Q, Yang X, Yu J, Palow C (2014) Uncovering large groups of active malicious accounts in online social networks. In: Proceedings of the ACM SIGSAC conference on computer and communications security, pp 477–488. doi:10.1145/2660267.2660269
8. Viswanath B, Bashir MA, Crovella M, Guha S, Gummadi KP, Krishnamurthy B, Mislove A (2014) Towards detecting anomalous user behavior in online social networks. In: Proceedings of the 23rd USENIX security symposium (USENIX security)
9. Fong S, Zhuang Y, He J (2012) Not every friend on a social network can be trusted: Classifying imposters using decision trees. In: Proceedings of the international conference on future generation communication technology (FGCT), pp 58–63 doi:10.1109/FGCT.2012.6476584
10. Stringhini G, Kruegel C, Vigna G (2010) Detecting spammers on social networks. In: Proceedings of the 26th annual computer security applications conference, pp 1–9. doi:10.1145/1920261.1920263
11. Verma M, Divya Sofat S (2014) Techniques to detect spammers in Twitter a survey. Int J Comput Appl 85(10):27–32
12. Everett RM, Nurse JRC, Erola A (2016) The anatomy of online deception: what makes automated text convincing? In: Proceedings of the 31st annual ACM symposium on applied computing. ACM, pp 1115–1120. doi:10.1145/2851613.2851813
13. Creese S, Goldsmith M, Nurse JRC, Phillips E (2012) A data-reachability model for elucidating privacy and security risks related to the use of online social networks. In: Proceedings of the 11th IEEE international conference on trust, security and privacy in computing and communications (TrustCom). IEEE, pp 1124–1131. doi:10.1109/TrustCom.2012.22
14. Bruce J, Scholtz J, Hodges D, Emanuel L, Fraser DS, Creese S, Love OJ (2014) Pathways to identity: using visualization to aid law enforcement in identification tasks. Secur Inf 3(1):1–13. doi:10.1186/s13388-014-0012-6
15. Pham DT, Dimov SS, Nguyen C (2005) Selection of k in k-means clustering. Proc Inst Mech Eng Part C J Mech Eng Sci 219(1):103–119
16. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
17. Jackson JE (2005) A user's guide to principal components. Wiley, New York